Introduction

Football is the most popular sport around the world with over three billion fans, and with professional players from almost every country. Today most countries have a professional league and have national teams competing to be a part of the world cup (The Sporting Blog, 2024). In this matter, football clubs must take big decisions when it comes to the selling and buying of players to their teams. These decisions can have crucial implications on their performance in the leagues and are thereby an interesting and relevant topic to study.

This research will have its focus on determining the most influential factors when determining the market values of professional soccer players using the website 'Transfermarkt' (Top Transfers, n.d.). The market value is an estimated value that the clubs can expect to pay for a particular player. The calculation of the values relies on the wisdom of the community of Transfermarkt, but it would be interesting to look into the features they might take into account.

By applying web scraping methods, it is possible to scrape the top transfers listed in Transfermarkt. Different player information will be extracted together with the market value, to conduct a comprehensive analysis of the features involved when determining the market value of a football player. This includes their position on the field, the age and the club and league the play in, together with the market value.

Research question and hypothesis

The research question for this paper will revolve around,

'Which features have the most effect on predicting the market value of a soccer player?'

The hypothesis that I expect is that the gathered features will have an effect on the prediction of market value, where especially the club they play will have the greatest impact. The reason for this hypothesis is the huge difference in the clubs' revenues. On the top 20 richest clubs the difference between first and twentieth is more than 400 million pounds, thereby players playing in some of the richest clubs might also be worth more. Further, is it also known that especially Premier League clubs have some of the richest clubs, and thereby it is also expected that the league feature will have an effect on the market value prediction (Goal, 2023)

Data gathering

To gather the necessary data about the players, their market values and the four different feature, age, club, league, and position, I will make use of web scraping techniques learned in class. All the data are scraped from the top transfer list from Transfermarkt (*Top Transfers*, n.d.). This list is a compilation of players from different leagues and clubs around the world and thereby an easy access to various professional players, which I can test the model and hypothesis on to answer the research question.

To create a predictive model, we need to extract the target variable and predictor variables. The target variable is the variable we want to predict using predictive modelling. In this research it will be the market value that is the target variable. The predictor variables are the input features that will be used to predict the target variable. I have selected four different features that I thought could be of importance in predicting the market value.

The first one is the players' age, as it is my idea that the older players especially those close to or above 30 are not athletically the same as the younger players. Further, is it also my thought that the younger players close to 20 do not have the same experience and are still not as hyped or as highly rated as the ones that have played in the professional leagues for a longer time. The second is the position they play. This feature I am little more unsure about, and actually do not think the correlation with market value is that big as all the positions are important for the team's game. However, I do think that the combination of position and age might have a correlation, as I know that e.g. goalkeepers usually can play until they are a bit older than the rest of the positions.

Lastly the third and fourth feature are the club and league that they are currently playing in. I think these two features have a huge impact on market value as earlier explained there is a big difference in how rich the teams are but also the leagues. Therefore, it can be the thought that a rich club also can afford players that are worth more and might also raise the players' market value that way.

Webscraping and Data cleaning

I use the Beautifulsoup library to webscrape from Transfermarkt. By clicking 'View Page Source' the HTML source of the page is available and we can easily search for the different features we are interested in. First step is to identify where in the HTML the features are located. Secondly, we need to create a code that can save the values in a list, so that we at last can create a csv file with all the players their target and predictor variables.

The web scraping code is built up by a for loop going through all of the pages. Each page consists of 25 players and there are 80 pages on the top transfer list. We can thereby retrieve 2000 entries/players. For each page relevant code is created to retrieve the necessary features and target

data. All the data is stored in lists named after each feature, and we can hereafter create a Pandas data frame and convert it to a csv file, that we can create our model on.

Now that we have all the data in a csv file it is time to prepare the data, so that we can use it to create a predictive model. The two three features positions, club and league are all string data and should be converted to float data. From sklearn preprocessing library, we can use labelencoder to convert the data into float. Furthermore if there is missing values, the value should be either replaced by the correct value, or the entry should be deleted. I choose to drop all the N/A as this only applied to 10 out of 2000 entries. Lastly was it interesting looking a possible outlier. In the picture below we can spot a few possible outliers, however especially on is far out from the average, as seen in the picture below. Thus, I remove this entry (Jude Bellingham), but the rest I will remain as they are not too far from the averages.

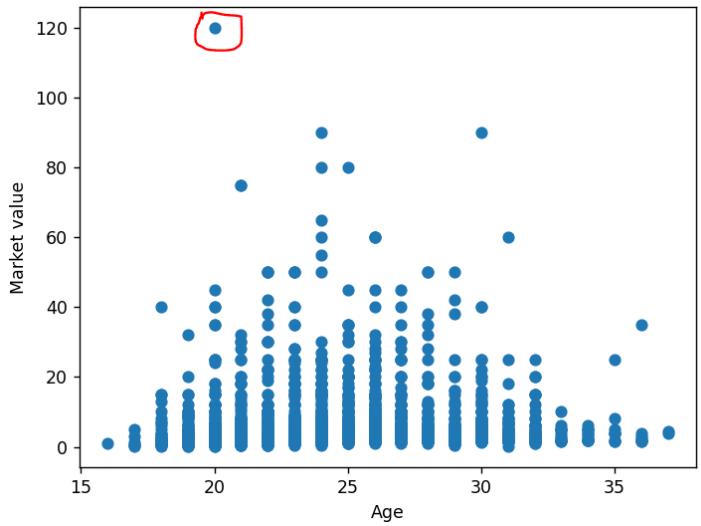

Figure 1: Scatter plot of age and market value, with Jude Bellingham marked.

Exploratory Data Analysis

For this section, I will make use of different plots teached in class. These will explore the data, before making a predictive model. Below can be seen two different histograms, that each represents the average market value for each league and for each position. It is obvious that especially with the leagues there are a huge difference in the average market value. Especially the English football league

(Premier League) has an average market value above 17,5 million. But also, the Saudi Pro League almost have an average of 12,5 million. However, most of them are just above or below 2,5 million.

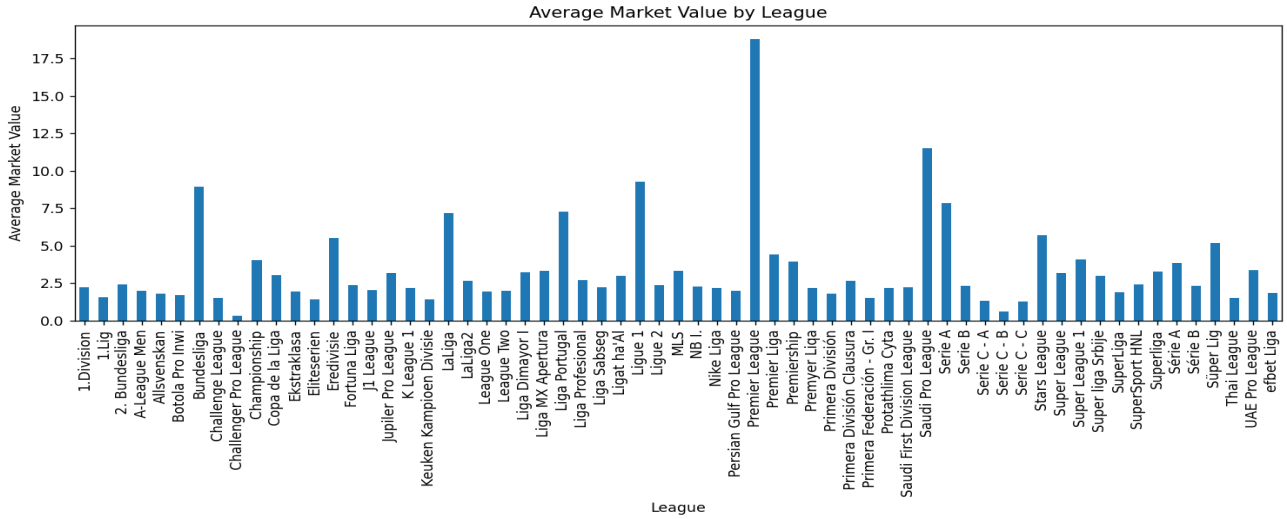

Figure 2: Histogram of Leagues and average market value

For the position histogram, there is not as much of a difference in the average market value. The lowest is the 'Left Midfield' just below 5 million, and the highest is the Second Striker just above 8 million.

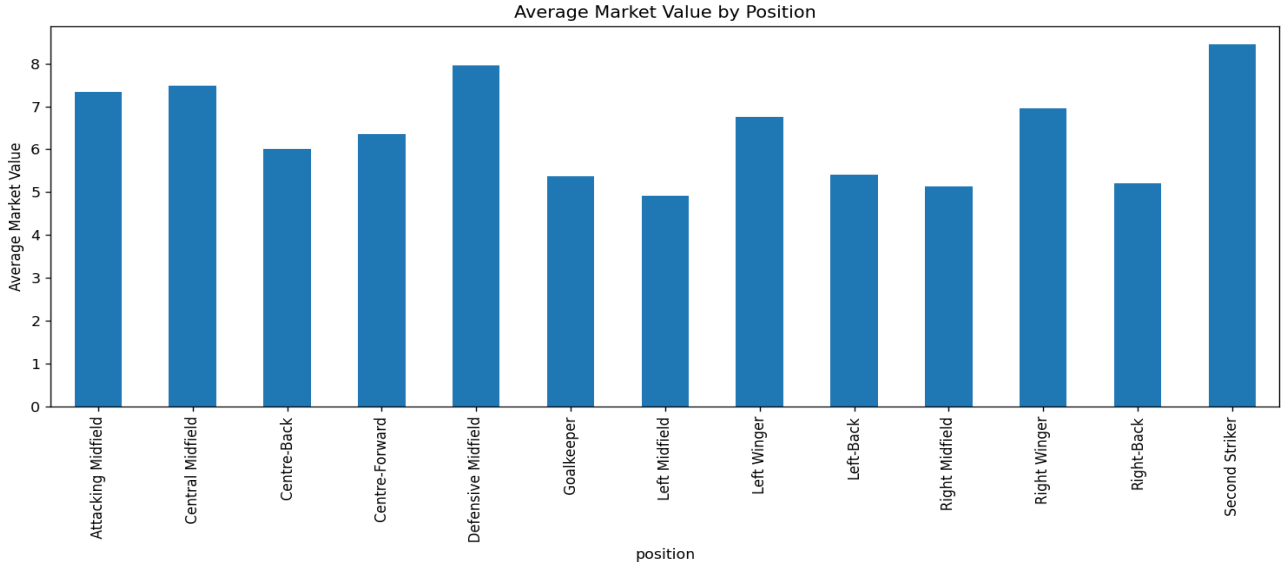

Figure 3: Histogram of positions and average market value

The graph and scatter plot below is similar to the one we saw earlier over the age and market value. We can easily detect the age peak compared to how much market value the player has. Players can be worth the most when they are around 24-25 of age, and especially below 20 and above 30 is when they have the least market value. This is also what could be expected as younger players have not yet created themselves a reputation, and are still in the age, where the player is being formed with a lot of practice to do. Opposite is it for the older players that are lacking towards the end of their soccer career, and therefore when a club buys the player it will be with the knowledge that they will not be able to sell the player for the same amount just a year later.

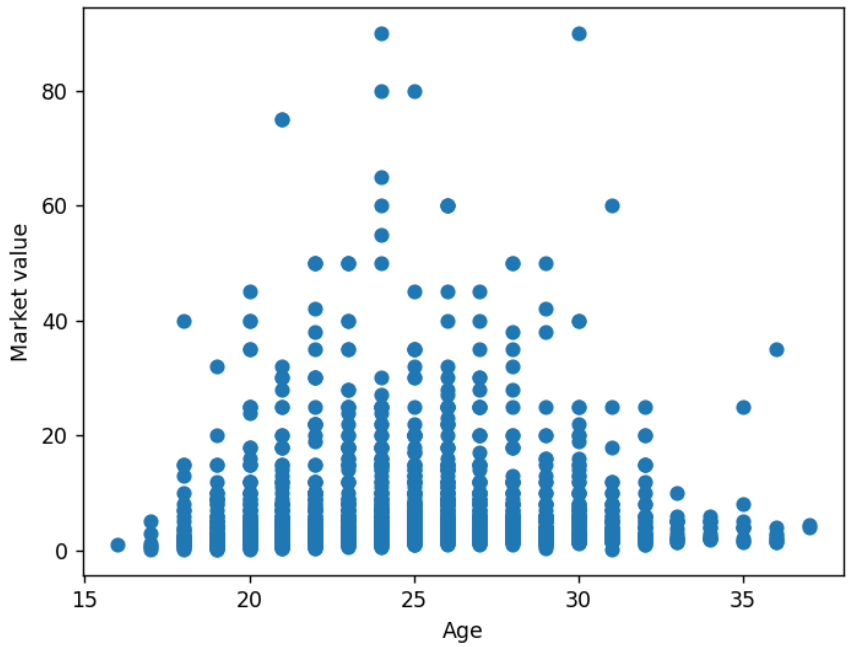

Figure 4: Scatter plot of age and market value.

Modelling

Firstly, when creating a machine learning model, it is important to identify weather we should make use of regression or classification. Classification models are valuable to use when predicting values which can be classified this could be e.g. male or female or a specific position on the soccer field. However, in this research the value being predicted are the market value of a player which is a continuous variable. and hence I will make use of a regression model,

Secondly, we need to decide on a regression model. Different regression models are amongst others linear regression, but also the Random Forest Regressor. Last option will create the base for the model in this research. This is a model that usually outperforms other regression models, especially when the correlation is nonlinear or interactions (Hindman, 2015). This is exactly the case here, where three of the features are categorized and the fourth does not show any linear correlation (see again figure… of Age and market value correlation). Further, is it my idea that the features might interact. An example is that goalkeepers usually play until they are older than the average player. Therefore, if you are a goalkeeper the optimal age for a high market value might be higher than the optimal for the rest of the positions. The random forest regressor is a working by dividing the feature data in two different subsets, where the randomness ensures no overfitting. Each node has the best predicting variable and continues with repeatedly splitting the data and aiming at minimizing the prediction errors. It stops as soon as there is no improvement in the predictions, this is a leaf node. Multiple trees are constructed and will together make the final predictions. It means the more trees the less variance but also more accuracy of the predictions (Hindman, 2015).

I am now ready to start the coding of my model. I have my CSV file with the scraped data, and I start by dividing my it into feature and target. As features I have the before mentioned, age, position, club, and league, where my target is the market value. Having it divided I turn the data into training and test data, this is what we call supervised learning models. The idea is to divide the data

in the two groups, where the test set can be used to judge the model's performance (Hindman, 2015). I chose 20% of the data to be for the test, a X_test with the features and a y_test with the target. Similarly, I have a X-train and a y_train. I use these four variables to start coding my random forest regression model, using a library from Python, where it takes the training data for training the model, and predicting the test target form the test features to check how well it performs.

I further improved the model performance by fine-tuning the hyperparameters. I made a grid of five hyper parameters, n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. These parameters respectively adjusts the number of trees in the forest, the max level in each tree, min of data point in the node before it splits the node further, min number of datapoint allowed in a leaf node, and the max features considered for splitting the node (Koehrsen, 2019). This grid was used for creating a randomized search cross-validation. This search tests a number of different combinations and cross validations, where my model creates 100 iterations with five folds. The number of iterations determines the width of the search space, where the number of folds reduce overfitting, however increasing it further would also increase the run time, which at the moment already takes several minutes (Koehrsen, 2019).

Results

Different metrics can be applied to evaluate the model and gain an insight and idea about how well the model performs. I have used the r2 score together with the root mean squared error (RMSE), and get the following results:

r2 score: 0.26489164503114937

Root Mean Squared Error: 7.003258698644453

Both metrices are indicators of how well a model performs. The r2 score explains how well the variables in the regression model can predict the target (Chugh, 2024). In this case the features are able to predict what is equivalent to 26,5%. In addition, the RMSE measures the standard deviation, the difference between the original and predicted value. In this case the RMSE is 7, which means that on average there is a deviation of 7 million (Chugh, 2024). Both numbers does not indicate a great model, however it seems that there actually is a correlation between the chosen feature and the market value, so a further investigation of how important each feature are and which feature is the most important will be interesting to analyze.

For this analysis I made use of the method feature_importances_, which tells which feature(s) are the most important when predicting our target (Kumar, 2023). The score in terms of random forest, is calculated on how often it is used in the leaf nodes and how much the contribution is (Kumar, 2023). The following scores are the feature importance of this regression model.

club 0.408595

league 0.270703

Age 0.175272

position 0.145430

Underneath is a visualization of the results in the form of a box plot model. What can be interpreted from these results is that the club and league have the highest importance in this model. Whereas, the

position has the least importance, which was also the idea from the exploratory analysis, where it showed little difference between the positions and their market value.

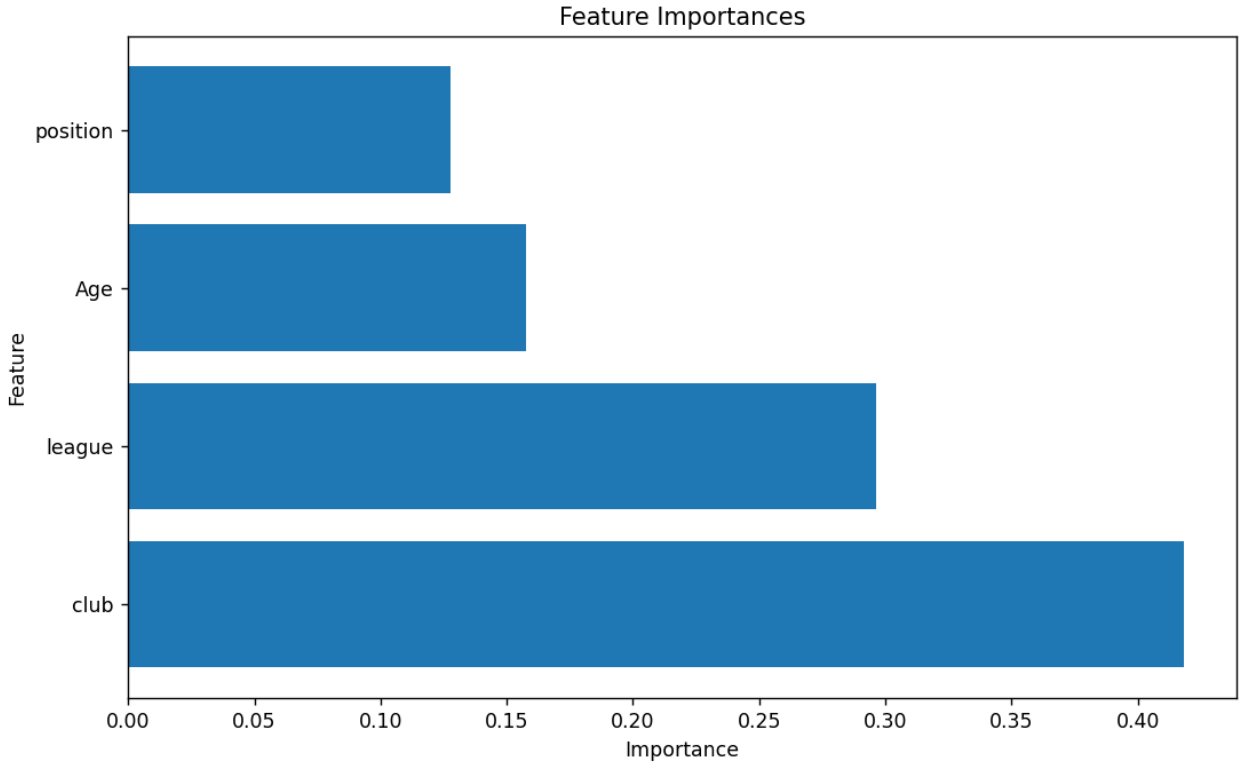

Figure 5: Box plot of the feature importance.

Conclusion

This research has extracted information on 2000 soccer players from the top transfer list on Tranfermarkt. The four features, age, position, club and league were scraped for all of the players together with their market value, in order to create a Random Forest regression model. The model turned out to show some correlation between the extracted features and the market value. However, with a r2 score of 26,5 and a RMSE on 7, it still indicated that a lot of improvement could be made on the model, for it to predict more precise the market value of a soccer player.

Nevertheless, feature importance could be applied to the model to gain an indication on how important the four features were. As a conclusion, we have found that the clubs are the most important

feature in determining the market value of soccer players, which was also the preliminary hypothesis. The club feature had a score of 40,9%, where the league had a score of 27,1%. However surprisingly, the age did not seem to have a high correlation, with a score on only 17,5%. The age feature seemed to show more correlation from the exploratory analysis (figure 4). Lastly, the feature with least importance was the position with only 14,5%, the low correlation was indicated by the exploratory analysis (figure 3), where the average market value did not change a lot according to the position.

Future research

Further research would be applying different machine learning models to find the best suited model for this specific research or even combining the models as Hindman (2015) suggests. Further would it be interesting to retrieve more data to make the model even better, where this research retrieved 2000 entries of the top transfer list future research could benefit from retrieving even more, making sure all the top leagues from around the world are equally represented.

Further, would it be beneficial to retrieve more features to predict the model on. It was obvious from the results of a r2 score of 0,265 That the features could predict some of the market value, but where still lacking a high percentage. A feature that could be worthy of note, would be the players' performances. This is a hard feature to extract, but player statistics such as number of goals or assists could be valuable for especially the performance of attacker positions. Other websites also have different ratings of the players that can be extracted. Lastly, can a feature such as previous clubs, and previous transfer fees also have an important effect on the market value.

References

Chugh, A. (2024, January 18). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e#:~:text=Both%20RMSE%20and%20R%2D%20Squared,variation%20in%20the%20response%20variable.

Goal (2023, December 21). *Which are the world's richest football clubs in 2023?* https://www.goal.com/en-za/news/which-are-the-worlds-richest-football-clubs-in-2021/psbb7gblbm6j1m5mc753tv1us

Hindman, M. (2015). "Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences". The ANNALS of the American Academy of Political and Social Science, 659(1), 48–62.

Koehrsen, W. (2019, December 10). *Hyperparameter Tuning the Random Forest in Python - Towards Data Science*. Medium. https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Kumar, A. (2023, December 9). *Feature Importance & Random Forest - Sklearn Python Example*. Analytics Yogi. https://vitalflux.com/feature-importance-random-forest-classifier-python/

The Sporting Blog (2024, January 3). *The 10 Most Popular Sports in the World: By Number of Fans and Participants — The Sporting Blog*. The Sporting Blog. https://thesporting.blog/blog/the-most-popular-sports-in-the-world

Top transfers. (n.d.). Transfermarkt. https://www.transfermarkt.com/transfers/saisontransfers/statistik?land_id=0&ausrichtung=&spielerposition_id=&altersklasse=&leihe=&transferfenster=&saison-id=0